Spectral Norm of Random Kernel Matrices with Applications to Privacy

Shiva Prasad Kasiviswanathan[1] **and Mark Rudelson**[2]

1 **Samsung Research America**
Mountain View, CA, USA
kasivisw@gmail.com
2 **Department of Mathematics, University of Michigan**
Ann Arbor, MI, USA
rudelson@umich.edu

Abstract

Kernel methods are an extremely popular set of techniques used for many important machine learning and data analysis applications. In addition to having good practical performance, these methods are supported by a well-developed theory. Kernel methods use an implicit mapping of the input data into a high dimensional feature space defined by a kernel function, i.e., a function returning the inner product between the images of two data points in the feature space. Central to any kernel method is the kernel matrix, which is built by evaluating the kernel function on a given sample dataset.

In this paper, we initiate the study of non-asymptotic spectral properties of random kernel matrices. These are $n \times n$ random matrices whose (i, j)th entry is obtained by evaluating the kernel function on $\mathbf{x}_i$ and $\mathbf{x}_j$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a set of n independent random high-dimensional vectors. Our main contribution is to obtain tight upper bounds on the spectral norm (largest eigenvalue) of random kernel matrices constructed by using common kernel functions such as polynomials and Gaussian radial basis.

As an application of these results, we provide lower bounds on the distortion needed for releasing the coefficients of kernel ridge regression under attribute privacy, a general privacy notion which captures a large class of privacy definitions. Kernel ridge regression is standard method for performing non-parametric regression that regularly outperforms traditional regression approaches in various domains. Our privacy distortion lower bounds are the first for any kernel technique, and our analysis assumes realistic scenarios for the input, unlike all previous lower bounds for other release problems which only hold under very restrictive input settings.

1998 ACM Subject Classification F. Theory of Computation

Keywords and phrases Random Kernel Matrices, Spectral Norm, Subguassian Distribution, Data Privacy, Reconstruction Attacks

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2015.898

1 Introduction

In recent years there has been significant progress in the development and application of kernel methods for many practical machine learning and data analysis problems. Kernel methods are regularly used for a range of problems such as classification (binary/multiclass), regression, ranking, and unsupervised learning, where they are known to almost always outperform "traditional" statistical techniques [23, 24]. At the heart of kernel methods is the notion of *kernel function*, which is a real-valued function of two variables. The power of kernel methods stems from the fact for every (positive definite) kernel function

18th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'15) / 19th Int'l Workshop on Randomization and Computation (RANDOM'15).
Editors: Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim; pp. 898–914

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

it is possible to define an inner-product and a lifting (which could be nonlinear) such that inner-product between any two lifted datapoints can be quickly computed using the kernel function evaluated at those two datapoints. This allows for introduction of nonlinearity into the traditional optimization problems (such as Ridge Regression, Support Vector Machines, Principal Component Analysis) without unduly complicating them.

The main ingredient of any kernel method is the *kernel matrix*, which is built using the kernel function, evaluated at given sample points. Formally, given a kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a sample set $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the kernel matrix K is an $n \times n$ matrix with its (i,j)th entry $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. Common choices of kernel functions include the polynomial kernel ($\kappa(\mathbf{x}_i, \mathbf{x}_j) = (a\langle \mathbf{x}_i, \mathbf{x}_j \rangle + b)^p$, for $p \in \mathbb{N}$) and the Gaussian kernel ($\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, for $a > 0$) [23, 24].

In this paper, we initiate the study of non-asymptotic spectral properties of *random kernel matrices*. A random kernel matrix, for a kernel function κ, is the kernel matrix K formed by n independent random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$. The prior work on random kernel matrices [13, 2, 6] have established various interesting properties of the spectral distributions of these matrices in the asymptotic sense (as $n, d \to \infty$). However, analyzing algorithms based on kernel methods typically requires understanding of the spectral properties of these random kernel matrices for *large, but fixed* n, d. A similar parallel also holds in the study of the spectral properties of "traditional" random matrices, where recent developments in the non-asymptotic theory of random matrices have complemented the classical random matrix theory that was mostly focused on asymptotic spectral properties [27, 20].

We investigate upper bounds on the largest eigenvalue (spectral norm) of random kernel matrices for polynomial and Gaussian kernels. We show that for inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn independently from a wide class of probability distributions over $\mathbb{R}^d$ (satisfying the subgaussian property), the spectral norm of a random kernel matrix constructed using a polynomial kernel of degree p, with high probability, is roughly bounded by $O(d^p n)$. In a similar setting, we show that the spectral norm of a random kernel matrix constructed using a Gaussian kernel is bounded by $O(n)$, and with high probability, this bound reduces to $O(1)$ under some stronger assumptions on the subgaussian distributions. These bounds are almost tight. Since the entries of a random kernel matrix are highly correlated, the existing techniques prevalent in random matrix theory cannot be directly applied. We overcome this problem by careful splitting and conditioning arguments on the random kernel matrix. Combining these with subgaussian norm concentrations form the basis of our proofs.

1.1 Applications

Largest eigenvalue of kernel matrices plays an important role in the analysis of many machine learning algorithms. Some examples include, bounding the Rademacher complexity for multiple kernel learning [16], analyzing the convergence rate of conjugate gradient technique for matrix-valued kernel learning [26], and establishing the concentration bounds for eigenvalues of kernel matrices [12, 25].

In this paper, we focus on an application of these eigenvalue bounds to an important problem arising while analyzing sensitive data. Consider a curator who manages a database of sensitive information but wants to release statistics about how a *sensitive* attribute (say, disease) in the database relates with some *nonsensitive* attributes (e.g., postal code, age, gender, etc). This setting is widely considered in the applied data privacy literature, partly since it arises with medical and retail data. Ridge regression is a well-known approach for solving these problems due to its good generalization performance. Kernel ridge regression is a powerful technique for building nonlinear regression models that operate by combining

ridge regression with kernel methods [21].[1] We present a *linear reconstruction attack* that reconstructs, with high probability, almost all the sensitive attribute entries given sufficiently accurate approximation of the kernel ridge regression coefficients. In a linear reconstruction attack, given the released information ρ, the attacker constructs a system of approximate linear equalities of the form $A\mathbf{z} \approx \rho$ for a matrix A and attempts to solve for $\mathbf{z}$.

We consider reconstruction attacks against *attribute privacy*, a loose notion of privacy, where the goal is to just avoid any gross violation of privacy. Concretely, the input is assumed to be a database whose ith row (record for individual i) is $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i \in \mathbb{R}^d$ is assumed to be known to the attacker (public information) and $y_i \in \{0, 1\}$ is the sensitive attribute, and a privacy mechanism is *attribute non-private* if the attacker can consistently reconstruct a large fraction of the sensitive attribute $(y_1, \ldots, y_n)$. We show that any privacy mechanism that always adds $\approx o(1/(d^p n))$ noise[2] to each coefficient of a polynomial kernel ridge regression model is attribute non-private. Similarly any privacy mechanism that always adds $\approx o(1)$ noise[2] to each coefficient of a Gaussian kernel ridge regression model is attribute non-private. As we later discuss, there exists natural settings of inputs under which these kernel ridge regression coefficients, even without the privacy constraint, have the same magnitude as these noise bounds, implying that privacy comes at a steep price. While the linear reconstruction attacks employed in this paper themselves are well-known [9, 15, 14], these are the first attribute privacy lower bounds that: (i) are applicable to any kernel method and (ii) work for any d-dimensional data, analyses of all previous attacks (for other release problems) require d to be comparable to n. Additionally, unlike previous reconstruction attack analyses, our bounds hold for a wide class of realistic distributional assumptions on the data.

1.2 Comparison with Related Work

In this paper, we study the largest eigenvalue of an $n \times n$ random kernel matrix in the non-asymptotic sense. The general goal with studying non-asymptotic theory of random matrices is to understand the spectral properties of random matrices, which are valid with high probability for matrices of a large fixed size. This is contrast with the existing theory on random kernel matrices which have focused on the asymptotics of various spectral characteristics of these random matrices, when the dimensions of the matrices tend to infinity. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be n i.i.d. random vectors. For any $F : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, symmetric in the first two variables, consider the random kernel matrix K with (i, j)th entry $K_{ij} = F(\mathbf{x}_i, \mathbf{x}_j, d)$. El Karoui [13] considered the case where K is generated by either the *inner-product kernels* (i.e., $F(\mathbf{x}_i, \mathbf{x}_j, d) = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle, d)$) or the *distance kernels* (i.e., $F(\mathbf{x}_i, \mathbf{x}_j, d) = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2, d)$). It was shown there that under some assumptions on f and on the distributions of $\mathbf{x}_i$'s, and in the "large d, large n" limit (i.e., and $d, n \to \infty$ and $d/n \to (0, \infty)$): a) the non-linear kernel matrix converges asymptotically in spectral norm to a linear kernel matrix, and b) there is a weak convergence of the limiting spectral density. These results were recently strengthened in different directions by Cheng *et al.* [2] and Do *et al.* [6]. To the best of our knowledge, ours is the first paper investigating the non-asymptotic spectral properties of a random kernel matrix.

Like the development of non-asymptotic theory of traditional random matrices has found multitude of applications in areas including statistics, geometric functional analysis, and compressed sensing [27], we believe that the growth of a non-asymptotic theory of random

[1] We provide a brief coverage of the basics of kernel ridge regression in Section 4.
[2] Ignoring the dependence on other parameters, including the regularization parameter of ridge regression.

kernel matrices will help in better understanding of many machine learning applications that utilize kernel techniques.

The goal of *private data analysis* is to release global, statistical properties of a database while protecting the privacy of the individuals whose information the database contains. Differential privacy [7] is a formal notion of privacy tailored to private data analysis. Differential privacy requires, roughly, that any single individual's data have little effect on the outcome of the analysis. A lot of recent research has gone in developing differentially private algorithms for various applications, including kernel methods [11]. A typical objective here is to release as accurate an approximation as possible to some function f evaluated on a database D.

In this paper, we follow a complementary line of work that seeks to understand how much distortion (noise) is necessary to privately release some particular function f evaluated on a database containing sensitive information [5, 8, 9, 15, 4, 18, 3, 19, 14]. The general idea here, is to provide *reconstruction attacks*, which are attacks that can reconstruct (almost all of) the sensitive part of database D given sufficiently accurate approximations to $f(D)$. Reconstruction attacks violate any *reasonable* notion of privacy (including, differential privacy), and the existence of these attacks directly translate into lower bounds on distortion needed for privacy.

Linear reconstruction attacks were first considered in the context of data privacy by Dinur and Nissim [5], who showed that any mechanism which answers $\approx n \log n$ random inner product queries on a database in $\{0,1\}^n$ with $o(\sqrt{n})$ noise per query is not private. Their attack was subsequently extended in various directions by [8, 9, 18, 3].

The results that are closest to our work are the attribute privacy lower bounds analyzed for releasing k-way marginals [15, 4], linear/logistic regression parameters [14], and a subclass of statistical M-estimators [14]. Kasiviswanathan *et al.* [15] showed that, if $d = \tilde{\Omega}(n^{1/(k-1)})$, then any mechanism which releases all k-way marginal tables with $o(\sqrt{n})$ noise per entry is attribute non-private.[3] These noise bounds were improved by De [4], who presented an attack that can tolerate a constant fraction of entries with arbitrarily high noise, as long as the remaining entries have $o(\sqrt{n})$ noise. Kasiviswanathan *et al.* [14] recently showed that, if $d = \Omega(n)$, then any mechanism which releases d different linear or logistic regression estimators each with $o(1/\sqrt{n})$ noise is attribute non-private. They also showed that this lower bound extends to a subclass of statistical M-estimator release problems. A point to observe is that in all the above referenced results, d has to be comparable to n, and this dependency looks unavoidable in those results due to their use of least singular value bounds. However, in this paper, our privacy lower bounds hold for all values of d, n (d could be $\ll n$). Additionally, all the previous reconstruction attack analyses critically require the $\mathbf{x}_i$'s to be drawn from product of univariate subgaussian distributions, whereas our analysis here holds for any d-dimensional subgaussian distributions (not necessarily product distributions), thereby is more widely applicable. The subgaussian assumption on the input data is quite common in the analysis of machine learning algorithms [1].

2 Preliminaries

2.1 Notation

We use $[n]$ to denote the set $\{1, \ldots, n\}$. $d_H(\cdot, \cdot)$ measures the Hamming distance. Vectors used in the paper are by default column vectors and are denoted by boldface letters. For

[3] The $\tilde{\Omega}$ notation hides polylogarithmic factors.

a vector $\mathbf{v}$, $\mathbf{v}^\top$ denotes its transpose and $\|\mathbf{v}\|$ denotes its Euclidean norm. For two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ denotes the inner product of $\mathbf{v}_1$ and $\mathbf{v}_2$. For a matrix M, $\|M\|$ denotes its spectral norm, $\|M\|_F$ denotes its Frobenius norm, and M_{ij} denotes its (i,j)th entry. $\mathbb{I}_n$ represents the identity matrix in dimension n. The unit sphere in d dimensions centered at origin is denoted by $S^{d-1} = \{\mathbf{z} : \|\mathbf{z}\| = 1, \mathbf{z} \in \mathbb{R}^d\}$. Throughout this paper C, c, C', also with subscripts, denote absolute constants (i.e., independent of d and n), whose value may change from line to line.

2.2 Background on Kernel Methods

We provide a very brief introduction to the theory of kernel methods; see the many books on the topic [23, 24] for further details.

▶ **Definition 1** (Kernel Function)**.** Let $\mathcal{X}$ be a non-empty set. Then a function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel function on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ over $\mathbb{R}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}}.$$

For any *symmetric and positive semidefinite*[4] kernel κ, by Mercer's theorem [17] there exists: (i) a unique functional Hilbert space $\mathcal{H}$ (referred to as the reproducing kernel Hilbert space, Definition 2) on $\mathcal{X}$ such that $\kappa(\cdot, \cdot)$ is the inner product in the space and (ii) a map ϕ defined as $\phi(\mathbf{x}) := \kappa(\cdot, \mathbf{x})$[5] that satisfies Definition 1. The function ϕ is called the *feature map* and the space $\mathcal{H}$ is called the *feature space*.

▶ **Definition 2** (Reproducing Kernel Hilbert Space)**.** A kernel $\kappa(\cdot, \cdot)$ is a reproducing kernel of a Hilbert space $\mathcal{H}$ if $\forall f \in \mathcal{H}$, $f(\mathbf{x}) = \langle \kappa(\cdot, \mathbf{x}), f(\cdot) \rangle_{\mathcal{H}}$. For a (compact) $\mathcal{X} \subseteq \mathbb{R}^d$, and a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, we say $\mathcal{H}$ is a Reproducing Kernel Hilbert Space if there $\exists \kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, s.t.: a) κ has the reproducing property, and b) κ spans $\mathcal{H} = \overline{\text{span}\{\kappa(\cdot, \mathbf{x}) : \mathbf{x} \in \mathcal{X}\}}$.

A standard idea used in the machine-learning community (commonly referred to as the "kernel trick") is that kernels allow for the computation of inner-products in high-dimensional feature spaces ($\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}}$) using simple functions defined on pairs of input patterns ($\kappa(\mathbf{x}, \mathbf{y})$), without knowing the ϕ mapping explicitly. This trick allows one to efficiently solve a variety of non-linear optimization problems. Note that there is no restriction on the dimension of the feature maps ($\phi(\mathbf{x})$), i.e., it could be of infinite dimension.

Polynomial and Gaussian are two popular kernel functions that are used in many machine learning and data mining tasks such as classification, regression, ranking, and structured prediction. Let the input space $\mathcal{X} = \mathbb{R}^d$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, these kernels are defined as:

1. **Polynomial Kernel:** $\kappa(\mathbf{x}, \mathbf{y}) = (a\langle \mathbf{x}, \mathbf{y} \rangle + b)^p$, with parameters $a, b \in \mathbb{R}$ and $p \in \mathbb{N}$. Here a is referred to as the slope parameter, $b \geq 0$ trades off the influence of higher-order versus lower-order terms in the polynomial, and p is the polynomial degree. For an input $\mathbf{x} \in \mathbb{R}^d$, the feature map $\phi(\mathbf{x})$ of the polynomial kernel is a vector with a polynomial in d number of dimensions [23].

[4] A positive definite kernel is a function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for any $n \geq 1$, for any finite set of points $\{\mathbf{x}_i\}_{i=1}^n$ in $\mathcal{X}$ and real numbers $\{a_i\}_{i=1}^n$, we have $\sum_{i,j=1}^n a_i a_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.

[5] $\kappa(\cdot, \mathbf{x})$ is a vector with entries $\kappa(\mathbf{x}', \mathbf{x})$ for all $\mathbf{x}' \in \mathcal{X}$.

2. **Gaussian Kernel:** (frequently referred to as the *radial basis kernel*): $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-a\|\mathbf{x}-\mathbf{y}\|^2\right)$ with real parameter $a > 0$. The value of a controls the locality of the kernel with low values indicating that the influence of a single point is "far" and vice-versa [23]. An equivalent popular formulation, is to set $a = 1/2\sigma^2$, and hence, $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2\right)$. For an input $\mathbf{x} \in \mathbb{R}^d$, the feature map $\phi(\mathbf{x})$ of the Gaussian kernel is a vector of infinite dimensions [23]. Note that while we focus on the Gaussian kernel in this paper, the extension of our results to other exponential kernels such as the *Laplacian kernel* (where $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-a\|\mathbf{x}-\mathbf{y}\|_1\right)$), is quite straightforward.

2.3 Background on Subgaussian Random Variables

Let us start by formally defining subgaussian random variables and vectors.

▶ **Definition 3** (Subgaussian Random Variable and Vector)**.** We call a random variable $x \in \mathbb{R}$ subgaussian if there exists a constant $C > 0$ if $\Pr[|x| > t] \leq 2\exp(-t^2/C^2)$ for all $t > 0$. We say that a random vector $\mathbf{x} \in \mathbb{R}^d$ is subgaussian if the one-dimensional marginals $\langle \mathbf{x}, \mathbf{y} \rangle$ are subgaussian random variables for all $\mathbf{y} \in \mathbb{R}^d$.

The class of subgaussian random variables includes many random variables that arise naturally in data analysis, such as standard normal, Bernoulli, spherical, bounded (where the random variable x satisfies $|x| \leq M$ *almost surely* for some fixed M). The natural generalizations of these random variables to higher dimension are all subgaussian random vectors. For many *isotropic convex sets*[6] $\mathcal{K}$ (such as the hypercube), a random vector $\mathbf{x}$ uniformly distributed in $\mathcal{K}$ is subgaussian.

▶ **Definition 4** (Norm of Subgaussian Random Variable and Vector)**.** The ψ_2-norm of a subgaussian random variable $x \in \mathbb{R}$, denoted by $\|x\|_{\psi_2}$ is:

$$\|x\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}[\exp(|x|^2/t^2)] \leq 2\right\}.$$

The ψ_2-norm of a subgaussian random vector $\mathbf{x} \in \mathbb{R}^d$ is:

$$\|\mathbf{x}\|_{\psi_2} = \sup_{\mathbf{y} \in S^{d-1}} \|\langle \mathbf{x}, \mathbf{y} \rangle\|_{\psi_2}.$$

▶ **Claim 5** (Vershynin [27])**.** *Let $x \in \mathbb{R}$ be a subgaussian random variable. Then there exists a constant $C > 0$, such that $\Pr[|x| > t] \leq 2\exp(-Ct^2/\|x\|_{\psi_2}^2)$.*

Consider a subset T of $\mathbb{R}^d$, and let $\epsilon > 0$. An ϵ-net of T is a subset $\mathcal{N} \subseteq T$ such that for every $\mathbf{x} \in T$, there exists a $\mathbf{z} \in \mathcal{N}$ such that $\|\mathbf{x} - \mathbf{z}\| \leq \epsilon$. We would use the following well-known result about the size of ϵ-nets.

▶ **Proposition 6** (Bounding the size of an ϵ-Net [27])**.** *Let T be a subset of S^{d-1} and let $\epsilon > 0$. Then there exists an ϵ-net of T of cardinality at most $(1 + 2/\epsilon)^d$.*

The proof of the following claim follows by standard techniques.

▶ **Claim 7** (Vershynin [27])**.** *Let $\mathcal{N}$ be a $1/2$-net of S^{d-1}. Then for any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\| \leq 2\max_{\mathbf{y} \in \mathcal{N}} \langle \mathbf{x}, \mathbf{y} \rangle$.*

[6] A convex set $\mathcal{K}$ in $\mathbb{R}^d$ is called isotropic if a random vector chosen uniformly from $\mathcal{K}$ according to the volume is isotropic. A random vector $\mathbf{x} \in \mathbb{R}^d$ is isotropic if for all $\mathbf{y} \in \mathbb{R}^d$, $\mathbb{E}[\langle \mathbf{x}, \mathbf{y} \rangle^2] = \|\mathbf{y}\|^2$.

3 Largest Eigenvalue of Random Kernel Matrices

In this section, we provide the upper bound on the largest eigenvalue of a random kernel matrix, constructed using polynomial or Gaussian kernels. Notice that the entries of a random kernel matrix are dependent. For example any triplet of entries (i, j), (j, k) and (k, i) are mutually dependent. Additionally, we deal with vectors drawn from general subgaussian distributions, and therefore, the coordinates within a random vector need not be independent.

We start off with a simple lemma, to bound the Euclidean norm of a subgaussian random vector. A random vector $\mathbf{x}$ is *centered* if $\mathbb{E}[\mathbf{x}] = 0$.

▶ **Lemma 8.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be independent centered subgaussian vectors. Then for all $i \in [n]$, $\Pr[\|\mathbf{x}_i\| \geq C\sqrt{d}] \leq \exp(-C'd)$ for constants C, C'.*

Proof. To this end, note that since $\mathbf{x}_i$ is a subgaussian vector (from Definition 3)

$$\Pr\left[|\langle \mathbf{x}_i, \mathbf{y}\rangle| \geq C\sqrt{d}/2\right] \leq 2\exp(-C_2 d),$$

for constants C and C_2, any unit vector $\mathbf{y} \in S^{d-1}$. Taking the union bound over a $(1/2)$-net $(\mathcal{N})$ in S^{d-1}, and using Proposition 6 for the size of the nets (which is at most 5^d as $\epsilon = 1/2$), we get that

$$\Pr\left[\max_{\mathbf{y}\in\mathcal{N}} |\langle \mathbf{x}_i, \mathbf{y}\rangle| \geq C\sqrt{d}/2\right] \leq \exp(-C_3 d),$$

From Claim 7, we know that $\|\mathbf{x}_i\| \leq 2\max_{\mathbf{y}\in\mathcal{N}} \langle \mathbf{x}_i, \mathbf{y}\rangle$. Hence, $\Pr\left[\|\mathbf{x}_i\| \geq C\sqrt{d}\right] \leq \exp(-C'd)$. ◀

3.1 Polynomial Kernel

We now establish the bound on the spectral norm of a polynomial kernel random matrix. We assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent vectors drawn according to a centered subgaussian distribution over $\mathbb{R}^d$. Let K_p denote the kernel matrix obtained using $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a polynomial kernel. Our idea to split the kernel matrix K_p into its diagonal and off-diagonal parts, and then bound the spectral norms of these two matrices separately. The diagonal part contains independent entries of the form $(a\|\mathbf{x}_i\|^2 + b)^p$, and we use Lemma 8 to bound its spectral norm. Dealing with the off-diagonal part of K_p is trickier because of the dependence between the entries, and here we bound the spectral norm by its Frobenius norm. We also verify the upper bounds provided in the following theorem by conducting numerical experiments (see Figure 1a).

▶ **Theorem 9.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be independent centered subgaussian vectors. Let $p \in \mathbb{N}$, and let K_p be the $n \times n$ matrix with (i, j)th entry $K_{p_{ij}} = (a\langle \mathbf{x}_i, \mathbf{x}_j\rangle + b)^p$. Assume that $n \leq \exp(C_1 d)$ for a constant C_1. Then there exists constants C_0, C_0' such that*

$$\Pr\left[\|K_p\| \geq C_0^p |a|^p d^p n + 2^{p+1}|b|^p n\right] \leq \exp(-C_0' d).$$

Proof. To prove the theorem, we split the kernel matrix K_p into the diagonal and off-diagonal parts. Let $K_p = D + W$, where D represents the diagonal part of K_p and W the off-diagonal part of K_p. Note that

$$\|K_p\| \leq \|D\| + \|W\| \leq \|D\| + \|W\|_F.$$

Let us estimate the norm of the diagonal part D first. From Lemma 8, we know that for all $i \in [n]$ with $C_3 = C'$,

$$\Pr\left[\|\mathbf{x}_i\| \geq C\sqrt{d}\right] = \Pr\left[\|\mathbf{x}_i\|^2 \geq (C\sqrt{d})^2\right] \leq \exp(-C_3 d).$$

Instead of $\|\mathbf{x}\|_i^2$, we are interested in bounding $(a\|\mathbf{x}_i\|^2 + b)^p$.

$$\Pr\left[\|\mathbf{x}_i\|^2 \geq (C\sqrt{d})^2\right] = \Pr\left[(a\|\mathbf{x}_i\|^2 + b)^p \geq (a(C\sqrt{d})^2 + b)^p\right]. \tag{1}$$

Consider $(a(C\sqrt{d})^2 + b)^p$. A simple inequality to bound $(a(C\sqrt{d})^2 + b)^p$ is[7]

$$(a(C\sqrt{d})^2 + b)^p \leq 2^p(|a|^p(C\sqrt{d})^{2p} + |b|^p).$$

Therefore,

$$\Pr\left[(a\|\mathbf{x}_i\|^2 + b)^p \geq 2^p(|a|^p(C\sqrt{d})^{2p} + |b|^p)\right] \leq \Pr\left[(a\|\mathbf{x}_i\|^2 + b)^p \geq (a(C\sqrt{d})^2 + b)^p\right].$$

Using (1) and substituting in the above equation, for any $i \in [n]$

$$\Pr\left[(a\|\mathbf{x}_i\|^2 + b)^p \geq 2^p(|a|^p C^{2p} d^p + |b|^p)\right] \leq \Pr\left[\|\mathbf{x}_i\| \geq C\sqrt{d}\right] \leq \exp(-C_3 d).$$

By applying a union bound over all n non-zero entries in D, we get that for all $i \in [n]$

$$\begin{aligned}\Pr\left[(a\|\mathbf{x}_i\|^2 + b)^p \geq 2^p(|a|^p C^{2p} d^p + |b|^p)\right] &\leq n \cdot \exp(-C_3 d) \leq \exp(C_1 d) \cdot \exp(-C_3 d) \\ &\leq \exp(-C_4 d),\end{aligned}$$

as we assumed that $n \leq \exp(C_1 d)$. This implies that

$$\Pr[\|D\| \geq 2^p(|a|^p C^{2p} d^p + |b|^p)] \leq \exp(-C_4 d). \tag{2}$$

We now bound the spectral norm of the off-diagonal part W using Frobenius norm as an upper bound on the spectral norm. Firstly note, by definition, for any $\mathbf{y} \in \mathbb{R}^d$, the random variable $\langle \mathbf{x}_i, \mathbf{y}\rangle$ is subgaussian with its ψ_2-norm at most $C_5\|\mathbf{y}\|$ for some constant C_5. This follows as:

$$\|\langle \mathbf{x}_i, \mathbf{y}\rangle\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}[\exp(\langle \mathbf{x}_i, \mathbf{y}\rangle^2/t^2)] \leq 2\right\} \leq C_5\|\mathbf{y}\|.$$

Therefore, for a fixed $\mathbf{x}_j$, $\|\langle \mathbf{x}_i, \mathbf{x}_j\rangle\|_{\psi_2} \leq C_5\|\mathbf{x}_j\|$. For $i \neq j$, conditioning on $\mathbf{x}_j$,

$$\Pr\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \geq \tau\right] = \mathbb{E}_{\mathbf{x}_j}\left[\Pr\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \geq \tau \mid \mathbf{x}_j\right]\right].$$

From Claim 5,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_j}\left[\Pr\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \geq \tau \mid \mathbf{x}_j\right]\right] \leq \mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_6\tau^2}{\|\langle \mathbf{x}_i, \mathbf{x}_j\rangle\|_{\psi_2}^2}\right)\right] &\leq \mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_6\tau^2}{(C_5\|\mathbf{x}_j\|)^2}\right)\right] \\ &= \mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_7\tau^2}{\|\mathbf{x}_j\|^2}\right)\right],\end{aligned}$$

[7] For any $a, b, m \in \mathbb{R}$ and $p \in \mathbb{N}$, $(a \cdot m + b)^p \leq 2^p(|a|^p|m|^p + |b|^p)$.

where the last inequality uses the fact that $\|\langle \mathbf{x}_i, \mathbf{x}_j\rangle\|_{\psi_2} \leq C_5\|\mathbf{x}_j\|$. Now let us condition the above expectation on the value of $\|\mathbf{x}_j\|$ based on whether $\|\mathbf{x}_j\| \geq C\sqrt{d}$ or $\|\mathbf{x}_j\| < C\sqrt{d}$. We can rewrite

$$\mathbb{E}_{\mathbf{x}_j}\left[\frac{-C_7\tau^2}{\|\mathbf{x}_j\|^2}\right] \leq \mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_7\tau^2}{C^2 d}\right) \;\middle|\; \|\mathbf{x}_j\| < C\sqrt{d}\right] \Pr[\|\mathbf{x}_j\| < C\sqrt{d}] + \mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_7\tau^2}{\|\mathbf{x}_j\|^2}\right) \;\middle|\; \|\mathbf{x}_j\| \geq C\sqrt{d}\right] \Pr[\|\mathbf{x}_j\| \geq C\sqrt{d}].$$

The above equation can be easily be simplified as:

$$\mathbb{E}_{\mathbf{x}_j}\left[\frac{-C_7\tau^2}{\|\mathbf{x}_j\|^2}\right] \leq \exp\left(\frac{-C_8\tau^2}{d}\right) + \mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_7\tau^2}{\|\mathbf{x}_j\|^2}\right) \;\middle|\; \|\mathbf{x}_j\| \geq C\sqrt{d}\right] \Pr[\|\mathbf{x}_j\| \geq C\sqrt{d}].$$

From Lemma 8, $\Pr[\|\mathbf{x}_j\| \geq C\sqrt{d}] \leq \exp(-C_3 d)$, and

$$\mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_7\tau^2}{\|\mathbf{x}_j\|^2}\right) \;\middle|\; \|\mathbf{x}_j\| \geq C\sqrt{d}\right] \leq 1.$$

This implies that as $\Pr[\|\mathbf{x}_j\| \geq C\sqrt{d}] \leq \exp(-C_3 d))$,

$$\mathbb{E}_{\mathbf{x}_j}\left[\exp\left(\frac{-C_7\tau^2}{\|\mathbf{x}_j\|^2}\right) \;\middle|\; \|\mathbf{x}_j\| \geq C\sqrt{d}\right] \Pr[\|\mathbf{x}_j\| \geq C\sqrt{d}] \leq \exp(-C_3 d).$$

Putting the above arguments together,

$$\Pr\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \geq \tau\right] = \mathbb{E}_{\mathbf{x}_j}\left[\Pr\left[|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \geq \tau \mid \mathbf{x}_j\right]\right] \leq \exp\left(\frac{-C_8\tau^2}{d}\right) + \exp(-C_3 d).$$

Taking a union bound over all $(n^2 - n) < n^2$ non-zero entries in W,

$$\Pr\left[\max_{i\neq j} |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \geq \tau\right] \leq n^2\left(\exp\left(\frac{-C_8\tau^2}{d}\right) + \exp(-C_3 d)\right).$$

Setting $\tau = C \cdot d$ in the above and using the fact that $n \leq \exp(C_1 d)$,

$$\Pr\left[\max_{i\neq j} |\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \geq C \cdot d\right] \leq \exp(-C_9 d). \tag{3}$$

We are now ready to bound the Frobenius norm of W.

$$\|W\|_F = \left(\sum_{i\neq j}(a\langle \mathbf{x}_i, \mathbf{x}_j\rangle + b)^{2p}\right)^{1/2} \leq \left(n^2 2^{2p}\left(|a|^{2p}\langle \mathbf{x}_i, \mathbf{x}_j\rangle^{2p} + |b|^{2p}\right)\right)^{1/2} \leq n2^p\left(|a|^p|\langle \mathbf{x}_i, \mathbf{x}_j\rangle|^p + |b|^p\right).$$

Plugging in the probabilistic bound on $|\langle \mathbf{x}_i, \mathbf{x}_j\rangle|$ from (3) gives,

$$\begin{aligned}\Pr\left[\|W\|_F \geq n2^p\left(|a|^p|C^p d^p + |b|^p\right)\right] &\leq \Pr\left[n2^p\left(|a|^p|\langle \mathbf{x}_i, \mathbf{x}_j\rangle|^p + |b|^p\right) \geq n2^p\left(|a|^p|C^p d^p + |b|^p\right)\right] \\ \Pr\left[\|W\|_F \geq n2^p\left(|a|^p|C^p d^p + |b|^p\right)\right] &\leq \Pr\left[n2^p\left(|a|^p|\langle \mathbf{x}_i, \mathbf{x}_j\rangle|^p + |b|^p\right) \geq n2^p\left(|a|^p|C^p d^p + |b|^p\right)\right] \\ &\leq \exp(-C_9 d). \end{aligned} \tag{4}$$

Plugging bounds on $\|D\|$ (from (2)) and $\|W\|_F$ (from (4)) to upper bound $\|K_p\| \leq \|D\| + \|W\|_F$ yields that there exists constants C_0 and C_0' such that,

$$\begin{aligned}\Pr\left[\|K_p\| \geq C_0^p|a|^p d^p n + 2^{p+1}|b|^p n\right] &\leq \Pr\left[\|D\| + \|W\|_F \geq C_0^p|a|^p d^p n + 2^{p+1}|b|^p n\right] \\ &\leq \exp(-C_0' d).\end{aligned}$$

This completes the proof of the theorem. The chain of constants can easily be estimated starting with the constant in the definition of the subgaussian random variable. ◀

▶ Remark. Note that for our proofs it is only necessary that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent random vectors, but they need not be identically distributed.

The above spectral norm upper bound on K_p (again with exponentially high probability) could be improved to

$$O\left(C_0^p|a|^p(d^p + d^{p/2}n) + 2^{p+1}n|b|^p\right),$$

with a slightly more involved analysis (omitted here). For an even p, the expectation of every individual entry of the matrix K_p is positive, which provides tight examples for this bound.

3.2 Gaussian Kernel

We now establish the bound on the spectral norm of a Gaussian kernel random matrix. Again assume $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent vectors drawn according to a centered subgaussian distribution over $\mathbb{R}^d$. Let K_g denote the kernel matrix obtained using $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a Gaussian kernel. Here an upper bound of n on the spectral norm on the kernel matrix follows trivially as all entries of K_g are less than equal to 1. We show that this bound is tight, in that for small values of a, with high probability the spectral norm is at least $\Omega(n)$ (Theorem 10 (Part 2)).

In fact, even for large a's, it is impossible to obtain better than $O(n)$ upper bound on the spectral norm of K_g without additional assumptions on the subgaussian distribution, as illustrated by this example: Consider a distribution over $\mathbb{R}^d$, such that a random vector drawn from this distribution is a zero vector $(0)^d$ with probability $1/2$ and uniformly distributed over the sphere in $\mathbb{R}^d$ of radius $2\sqrt{d}$ with probability $1/2$. A random vector $\mathbf{x}$ drawn from this distribution is isotropic and subgaussian, but $\Pr[\mathbf{x} = (0)^d] = 1/2$. Therefore, in $\mathbf{x}_1, \ldots, \mathbf{x}_n$ drawn from this distribution, with high probability more than a constant fraction of the vectors will be $(0)^d$. This means that a proportional number of entries of the matrix K_g will be 1, and the norm will be $O(n)$ regardless of a.

This situation changes, however, when we add the additional assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have independent centered subgaussian coordinates[8] (i.e., each $\mathbf{x}_i$ is drawn from a product distribution formed from some d centered univariate subgaussian distributions). In that case, the kernel matrix K_g is a small perturbation of the identity matrix, and we show that the spectral norm of K_g is with high probability bounded by an absolute constant (for $a = \Omega(\log n/d)$). For this proof, similar to Theorem 9, we split the kernel matrix into its diagonal and off-diagonal parts. The spectral norm of the off-diagonal part is again bounded by its Frobenius norm. We also verify the upper bounds presented in the following theorem by conducting numerical experiments (see Figure 1b).

[8] Some of the commonly used subgaussian random vectors such as the standard normal, Bernoulli satisfy this additional assumption.

▶ **Theorem 10.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be independent centered subgaussian vectors. Let $a > 0$, and let K_g be the $n \times n$ matrix with (i,j)th entry $K_{g_{ij}} = \exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. Then there exists constants c, c_0, c_0', c_1 such that*

1. *$\|K_g\| \leq n$.*
2. *If $a < c_1/d$, $\Pr[\|K_g\| \geq c_0 n] \geq 1 - \exp(-c_0' n)$.*
3. *If all the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ satisfy the additional assumption of having independent centered subgaussian coordinates, and assume $n \leq \exp(C_1 d)$ for a constant C_1. Then for any $\delta > 0$ and $a \geq (2+\delta)\frac{\log n}{d}$, $\Pr[\|K_g\| \geq 2] \leq \exp(-c\zeta^2 d)$ with $\zeta > 0$ depending only on δ.*

Proof. Proof of Part 1 is straightforward as all entries of K_g do not exceed 1.

Let us prove the lower estimate for the norm in Part 2. For $i = 1, \ldots, n$ define

$$Z_i = \sum_{j=\frac{n}{2}+1}^{n} K_{g_{ij}}.$$

From Lemma 8 for all $i \in [n]$, $\Pr\left[\|\mathbf{x}_i\| \geq C\sqrt{d}\right] \leq \exp(-C'd)$. In other words, $\|\mathbf{x}_i\|$ is less than $C\sqrt{d}$ for all $i \in [d]$ with probability at least $1 - \exp(-C'd)$. Let us call this event $\mathcal{E}_1$. Under $\mathcal{E}_1$ and assumption $a < c_1/d$, $\mathbb{E}[Z_i] \geq c_2 n$ and $\mathbb{E}[Z_i^2] \leq c_3 n^2$. Therefore, by Paley-Zygmund inequality (under event $\mathcal{E}_1$),

$$\Pr[Z_i \geq c_4 n] \geq c_5. \tag{5}$$

Now $Z_1, \ldots, Z_n$ are not independent random variables. But if we condition on $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n$, then $Z_1, \ldots, Z_{n/2}$ become independent (for simplicity, assume that n is divisible by 2). Thereafter, an application of Chernoff bound on $Z_1, \ldots, Z_{n/2}$ using the probability bound from (5) (under conditioning on $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n$ and event $\mathcal{E}_1$) gives:

$$\Pr\left[Z_i \geq c_4 n \text{ for at least } c_5 n \text{ entries } Z_i \in \{Z_1, \ldots, Z_{n/2}\}\right] \geq 1 - \exp(-c_6 n).$$

The first conditioning can be removed by taking the expectation with respect to $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n$ without disturbing the exponential probability bound. Similarly, conditioning on event $\mathcal{E}_1$ can also be easily removed.

Let K_g' be the submatrix of K_g consisting of rows $1 \leq i \leq n/2$ and columns $n/2+1 \leq j \leq n$. Note that $\|K_g'\| \geq \mathbf{u}^\top K_g' \mathbf{u}$, where $\mathbf{u} = \left(\sqrt{\frac{2}{n}}, \ldots, \sqrt{\frac{2}{n}}\right)$ (of dimension $n/2$). Then

$$\Pr[\|K_g\| \leq c_0 n] \leq \Pr[\|K_g'\| \leq c_7 n] \leq \Pr[\mathbf{u}^\top K_g' \mathbf{u} \leq c_7 n]$$
$$\Pr\left[\frac{2}{n}\sum_{i=1}^{n/2} Z_i \leq c_7 n\right] \leq \exp(-c_0' n).$$

The last line follows as from above arguments with exponentially high probability above more than $\Omega(n)$ entries in $Z_1, \ldots, Z_{n/2}$ are greater than $\Omega(n)$, and by readjusting the constants.

Proof of Part 3: As in Theorem 9, we split the matrix K_g into the diagonal (D) and the off-diagonal part (W) (i.e., $K_g = D + W$). It is simple to observe that $D = \mathbb{I}_n$, therefore we just concentrate on W. The (i,j)th entry in W is $\exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are independent vectors with independent centered subgaussian coordinates. Therefore, we can use Hoeffding's inequality, for fixed i, j,

$$\Pr\left[\exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2) \geq \exp(-a(1-\zeta)d)\right] = \Pr\left[\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{d} \leq (1-\zeta)\right] \leq \exp(-c_8\zeta^2 d), \tag{6}$$

where we used the fact that if a random variable is subgaussian then its square is a subexponential random variable [27].[9] To estimate the norm of W, we bound it by its Frobenius norm. If $a \geq (2+\delta)\frac{\log n}{d}$, then we can choose $\zeta > 0$ depending on δ such that $n^2 \exp(-a(1-\zeta)d) \leq 1$. Hence,

$$\begin{aligned}
\Pr[\|K_g\| \geq 2] &\leq \Pr[\|D\| + \|W\|_F \geq 2] = \Pr[\|W\|_F \geq 1] \\
&= \Pr\left[\sum_{1\leq i,j\leq n, i\neq j} \exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2) \geq 1\right] \\
&\leq \Pr\left[\sum_{1\leq i,j\leq n, i\neq j} \exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2) \geq n^2 \exp(-a(1-\zeta)d)\right] \\
&\leq \Pr\left[\sum_{1\leq i,j\leq n} \exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2) \geq n^2 \exp(-a(1-\zeta)d)\right] \\
&\leq n^2 \Pr\left[\max_{1\leq i,j\leq n} \exp(-a\|\mathbf{x}_i - \mathbf{x}_j\|^2) \geq \exp(-a(1-\zeta)d)\right] \\
&\leq n^2 \exp(-c_8\zeta^2 d) \\
&\leq \exp(-c\zeta^2 d) \text{ for some constant } c.
\end{aligned}$$

The first equality follows as $\|D\| = 1$, and the second-last inequality follows from (6). This completes the proof of the theorem. Again the long chain of constants can easily be estimated starting with the constant in the definition of the subgaussian random variable. ◀

▶ Remark. Note that again the $\mathbf{x}_i$'s need not be identically distributed.

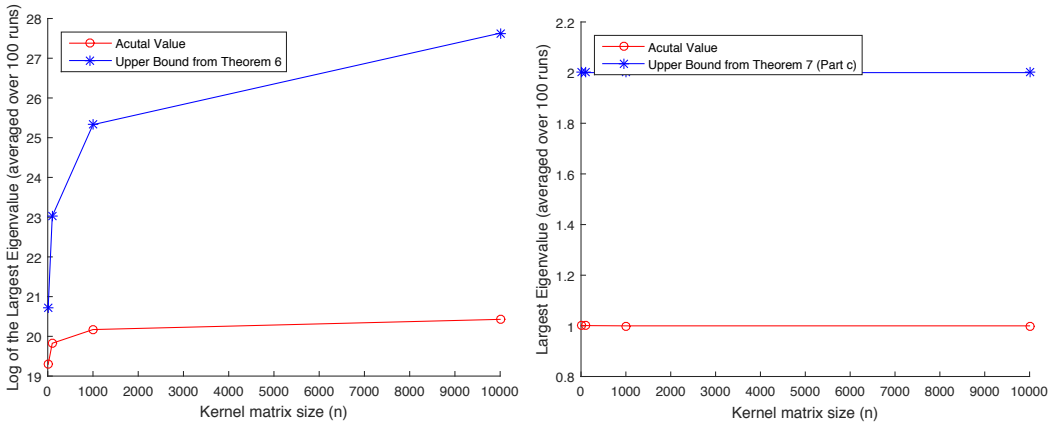
The analysis in Theorem 10 could easily be reworked to handle other exponential kernels such as the Laplacian kernel.

4 Privately Releasing Kernel Ridge Regression Coefficients

We consider an application of Theorems 9 and 10 to obtain noise lower bounds for privately releasing coefficients of kernel ridge regression. For privacy violation, we consider a generalization of *blatant non-privacy* [5] referred to as attribute non-privacy (formalized in [15]). Consider a database $D \in \mathbb{R}^{n\times d+1}$ that contains, for each individual i, a sensitive attribute $y_i \in \{0,1\}$ as well as some other information $\mathbf{x}_i \in \mathbb{R}^d$ which is assumed to be known to the attacker. The ith record is thus $(\mathbf{x}_i, y_i)$. Let $X \in \mathbb{R}^{n\times d}$ be a matrix whose ith row is $\mathbf{x}_i$, and let $\mathbf{y} = (y_1, \ldots, y_n)$. We denote the entire database $D = (X|\mathbf{y})$ where $|$ represents vertical concatenation. Given some released information ρ, the attacker constructs an estimate $\hat{\mathbf{y}}$ that she hopes is close to $\mathbf{y}$. We measure the attack's success in terms of the Hamming distance $d_H(\mathbf{y}, \hat{\mathbf{y}})$. A scheme is *not* attribute private if an attacker can consistently get an estimate that is within distance $o(n)$. Formally:

▶ **Definition 11** (Failure of Attribute Privacy [15])**.** A (randomized) mechanism $\mathcal{M} : \mathbb{R}^{n\times d+1} \to \mathbb{R}^l$ is said to allow (θ, γ)-attribute reconstruction if there exists a setting of the nonsensitive

[9] We call a random variable $x \in \mathbb{R}$ subexponential if there exists a constant $C > 0$ if $\Pr[|x| > t] \leq 2\exp(-t/C)$ for all $t > 0$.



(a) Polynomial Kernel **(b)** Gaussian Kernel

Figure 1 Largest eigenvalue distribution for random kernel matrices constructed with a polynomial kernel (left plot) and a Gaussian kernel (right plot). The actual value plots are constructed by averaging over 100 runs, and in each run we draw n independent standard Gaussian vectors in $d = 100$ dimensions. The predicted values are computed from bounds in Theorems 9 and 10 (Part 3). The kernel matrix size n is varied from 10 to 10000 in multiples of 10. For the polynomial kernel, we set $a = 1, b = 1$, and $p = 4$, and for the Gaussian kernel $a = 3\log(n)/d$. Note that our upper bounds are fairly close to the actual results. For the Gaussian kernel, the actual values are very close to 1.

attributes $X \in \mathbb{R}^{n\times d}$ and an algorithm (adversary) $\mathcal{A} : \mathbb{R}^{n\times d} \times \mathbb{R}^l \to \mathbb{R}^n$ such that for every $\mathbf{y} \in \{0,1\}^n$,

$$\Pr_{\rho \leftarrow \mathcal{M}((X|\mathbf{y}))}[\mathcal{A}(X,\rho) = \hat{\mathbf{y}} : d_H(\mathbf{y},\hat{\mathbf{y}}) \le \theta] \ge 1 - \gamma.$$

Asymptotically, we say that a mechanism is *attribute nonprivate* if there is an infinite sequence of n for which $\mathcal{M}$ allows $(o(n), o(1))$-attribute reconstruction. Here $d = d(n)$ is a function of n. We say the attack $\mathcal{A}$ is *efficient* if it runs in time $\text{poly}(n,d)$.

4.1 Kernel Ridge Regression Background

One of the most basic regression formulation is that of ridge regression [10]. Suppose that we are given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consisting of n points with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Here $\mathbf{x}_i$'s are referred to as the *regressors* and y_i's are the *response variables*. In linear regression the task is to find a linear function that models the dependencies between $\mathbf{x}_i$'s and the y_i's. A common way to prevent overfitting in linear regression is by adding a penalty regularization term (also known as *shrinkage* in statistics). In kernel ridge regression [21], we assume a model of form $y = f(\mathbf{x}) + \xi$, where we are trying to estimate the regression function f and ξ is some unknown vector that accounts for discrepancy between the actual response (y) and predicted outcome ($f(\mathbf{x})$). Given a reproducing kernel Hilbert space $\mathcal{H}$ with kernel κ, the goal of ridge regression kernel ridge regression is to estimate the unknown function $f^\star$ such the least-squares loss defined over the dataset with a weighted penalty based on the squared Hilbert norm is minimized.

$$\text{Kernel Ridge Regression:} \qquad \text{argmin}_{f\in\mathcal{H}} \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 + \lambda\|f\|_{\mathcal{H}}^2 \right), \tag{7}$$

where $\lambda > 0$ is a regularization parameter. By representer theorem [22], any solution $f^\star$ for (7), takes the form

$$f^\star(\cdot) = \sum_{i=1}^{n} \alpha_i \kappa(\cdot, \mathbf{x}_i), \tag{8}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)$ is known as the kernel ridge regression coefficient vector. Plugging this representation into (7) and solving the resulting optimization problem (in terms of α now), we get that the minimum value is achieved for $\alpha = \alpha^\star$, where

$$\alpha^\star = (K + \lambda \mathbb{I}_n)^{-1}\mathbf{y}, \text{ where } K \text{ is the kernel matrix with}$$
$$K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \text{ and } \mathbf{y} = (y_1, \ldots, y_n). \tag{9}$$

Plugging this $\alpha^\star$ from (9) in to (8), gives the final form for estimate $f^\star(\cdot)$. For a new point $\mathbf{x} \in \mathbb{R}^d$, the predicted response is $f^\star(\mathbf{x}) = \sum_{i=1}^n \alpha_i^\star \kappa(\mathbf{x}, \mathbf{x}_i)$ where $\alpha^\star = (K + \lambda \mathbb{I}_n)^{-1}\mathbf{y}$ and $\alpha^\star = (\alpha_1^\star, \ldots, \alpha_n^\star)$. Therefore, knowledge of $\alpha^\star$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ suffices for making future predictions.

If K is constructed using a polynomial kernel (defined in 1. in Section 2.2) then the above procedure is referred to as the *polynomial kernel ridge regression*, and similarly if K is constructed using a Gaussian kernel (defined in 2. in Section 2.2) then the above procedure is referred to as the *Gaussian kernel ridge regression*.

4.2 Reconstruction Attack from Noisy α^*

Algorithm 1 outlines the attack. The privacy mechanism releases a noisy approximation to $\alpha^\star$. Let $\tilde{\alpha}$ be this noisy approximation, i.e., $\tilde{\alpha} = \alpha^\star + \mathbf{e}$ where $\mathbf{e}$ is some unknown noise vector. The adversary tries to reconstruct an approximation $\hat{\mathbf{y}}$ of $\mathbf{y}$ from $\tilde{\alpha}$. The adversary solves the following ℓ_2-minimization problem to construct $\hat{\mathbf{y}}$:

$$\min_{\mathbf{z} \in \mathbb{R}^n} \|\tilde{\alpha} - (K + \lambda \mathbb{I}_n)^{-1}\mathbf{z}\|. \tag{10}$$

In the setting of attribute privacy, the database $D = (X|\mathbf{y})$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the rows of X, using which the adversary can construct K to carry out the attack. Since the matrix $K + \lambda \mathbb{I}_n$ is invertible for $\lambda > 0$ as K is a positive semidefinite matrix, the solution to (10) is simply $\mathbf{z} = (K + \lambda \mathbb{I}_n)\tilde{\alpha}$, element-wise rounding of which to closest $0, 1$ gives $\hat{\mathbf{y}}$.

Algorithm 1 Reconstruction Attack from Noisy Kernel Ridge Regression Coefficients

Input: Public information $X \in \mathbb{R}^{n \times d}$, regularization parameter λ, and $\tilde{\alpha}$ (noisy $\alpha^\star$).

1: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the rows of X, construct the kernel matrix K with $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$
2: **R**eturn $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ defined as follows:

$$\hat{y}_i = \begin{cases} 0 & \text{if } i\text{th entry in } (K + \lambda \mathbb{I}_n)\tilde{\alpha} < 1/2 \\ 1 & \text{otherwise} \end{cases}$$

▶ **Lemma 12.** *Let $\tilde{\alpha} = \alpha^\star + \mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^n$ is some unknown (noise) vector. If $\|\mathbf{e}\|_\infty \leq \beta$ (absolute value of all entries in $\mathbf{e}$ is less than β), then $\hat{\mathbf{y}}$ returned by Algorithm 1 satisfies, $d_H(\mathbf{y}, \hat{\mathbf{y}}) \leq 4(K + \lambda)^2 \beta^2 n$. In particular, if $\beta = o\left(\frac{1}{\|K\| + \lambda}\right)$, then $d_H(\mathbf{y}, \hat{\mathbf{y}}) = o(n)$.*

Proof. Since $\alpha^\star = (K + \lambda \mathbb{I}_n)^{-1}\mathbf{y}$, $\tilde{\alpha} = (K + \lambda \mathbb{I}_n)^{-1}\mathbf{y} + \mathbf{e}$. Now multiplying $(K + \lambda \mathbb{I}_n)$ on both sides gives,

$$(K + \lambda \mathbb{I}_n)\tilde{\alpha} = \mathbf{y} + (K + \lambda \mathbb{I}_n)\mathbf{e}.$$

Concentrate on $\|(K + \lambda \mathbb{I}_n)\mathbf{e}\|$. This can be bound as

$$\|(K + \lambda \mathbb{I}_n)\mathbf{e}\| \leq \|(K + \lambda \mathbb{I}_n)\|\|\mathbf{e}\| = (\|K\| + \lambda)\|\mathbf{e}\|.$$

If the absolute value of all the entries in $\mathbf{e}$ are less than β then $\|\mathbf{e}\| \leq \beta\sqrt{n}$. A simple manipulation then shows that if the above hold then $(K + \lambda \mathbb{I}_n)\mathbf{e}$ cannot have more than $4(\|K\| + \lambda)^2\beta^2 n$ entries with absolute value above $1/2$. Since $\hat{\mathbf{y}}$ and $\mathbf{y}$ only differ in those entries where $(K + \lambda \mathbb{I}_n)\mathbf{e}$ is greater than $1/2$, it follows that $d_H(\mathbf{y}, \hat{\mathbf{y}}) \leq 4(\|K\| + \lambda)^2\beta^2 n$. Setting $\beta = o(\frac{1}{\|K\|+\lambda})$ implies $d_H(\mathbf{y}, \hat{\mathbf{y}}) = o(n)$. ◀

For a privacy mechanism to be attribute non-private, the adversary has to be able reconstruct an $1 - o(1)$ fraction of $\mathbf{y}$ with high probability. Using the above lemma, and the different bounds on $\|K\|$ established in Theorems 9 and 10, we get the following lower bounds for privately releasing kernel ridge regression coefficients.

▶ **Theorem 13.**

1. *Any privacy mechanism which for every database $D = (X|\mathbf{y})$ where $X \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \{0,1\}^n$ releases the coefficient vector of a polynomial kennel ridge regression model (for constants a, b, and p) fitted between X (matrix of regressor values) and $\mathbf{y}$ (response vector), by adding $o(\frac{1}{d^p n+\lambda})$ noise to each coordinate is attribute non-private. The attack that achieves this attribute privacy violation operates in $O(dn^2)$ time.*
2. *Any privacy mechanism which for every database $D = (X|\mathbf{y})$ where $X \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \{0,1\}^n$ releases the coefficient vector of a Gaussian kennel ridge regression model (for constant a) fitted between X (matrix of regressor values) and $\mathbf{y}$ (response vector), by adding $o(\frac{1}{2+\lambda})$ noise to each coordinate is attribute non-private. The attack that achieves this attribute privacy violation operates in $O(dn^2)$ time.*

Proof. For Part 1, draw each individual i's non-sensitive attribute vector $\mathbf{x}_i$ independently from any d-dimensional subgaussian distribution, and use Lemma 12 in conjunction with Theorem 9.

For Part 2, draw each individual i's non-sensitive attribute vector $\mathbf{x}_i$ independently from any product distribution formed from some d centered univariate subgaussian distributions, and use Lemma 12 in conjunction with Theorem 10 (Part 3).[10]

The time needed to construct the kernel matrix K is $O(dn^2)$, which dominates the overall computation time. ◀

We can ask how the above distortion needed for privacy compares to typical entries in $\alpha^\star$. The answer is not simple, but there are natural settings of inputs, where the noise needed for privacy becomes comparable with coordinates of $\alpha^\star$, implying that the privacy comes at a steep price. One such example is if the $\mathbf{x}_i$'s are drawn from the standard normal distribution, $\mathbf{y} = (1)^n$, and all other kernel parameters are constant, then the expected value of the corresponding $\alpha^\star$ coordinates match the noise bounds obtained in Theorem 13.

[10] Note that it is not critical for $\mathbf{x}_i$'s to be drawn from a product distribution. It is possible to analyze the attack even under a (weaker) assumption that each individual i's non-sensitive attribute vector $\mathbf{x}_i$ is drawn independently from a d-dimensional subgaussian distribution, by using Lemma 12 in conjunction with Theorem 10 (Part 1).

Note that Theorem 13 makes no assumptions on the dimension d of the data, and holds for all values of n, d. This is different from all other previous lower bounds for attribute privacy [15, 4, 14], all of which require d to be comparable to n, thereby holding only either when the non-sensitive data (the $\mathbf{x}_i$'s) are very high-dimensional or for very small n. Also all the previous lower bound analyses [15, 4, 14] critically rely on the fact that the individual coordinates of each of the $\mathbf{x}_i$'s are independent[11], which is not essential for Theorem 13.

4.3 Note on using ℓ_1-reconstruction Attacks

A natural alternative to (10) is to use ℓ_1-minimization (also known as "LP decoding"). This gives rise to the following linear program:

$$\min_{\mathbf{z}\in\mathbb{R}^n} \|\tilde{\alpha} - (K + \lambda \mathbb{I}_n)^{-1}\mathbf{z}\|_1. \tag{11}$$

In the context of privacy, the ℓ_1-minimization approach was first proposed by Dwork *et al.* [8], and recently reanalyzed in different contexts by [4, 14]. These results have shown that, for some settings, the ℓ_1-minimization can handle considerably more complex noise patterns than the ℓ_2-minimization. However, in our setting, since the solutions for (11) and (10) are exactly the same ($\mathbf{z} = (K + \lambda \mathbb{I}_n)\tilde{\alpha}$), there is no inherent advantage of using the ℓ_1-minimization.

5 Concluding Remarks

We initiate the study of non-asymptotic spectral properties of random kernel matrices, and provide tight bounds on the spectral norm of these matrices when constructed using kernel functions such as polynomials and Gaussian radial basis. Using these results, we provide lower bounds on the distortion needed for releasing coefficients of kernel ridge regression under attribute privacy, a general privacy notion that captures a large class of privacy definitions.

We believe that developing a non-asymptotic spectral theory for random kernel matrices is an interesting research direction that would provide deep insights into the workings of many kernel-based machine learning algorithms.

Acknowledgements. We are grateful for helpful initial discussions with Adam Smith and Ambuj Tewari.

References

1 Olivier Bousquet, Ulrike von Luxburg, and G Rätsch. Advanced Lectures on Machine Learning. In *ML Summer Schools 2003*, 2004.

2 Xiuyuan Cheng and Amit Singer. The Spectrum of Random Inner-Product Kernel Matrices. *Random Matrices: Theory and Applications*, 2(04), 2013.

3 Krzysztof Choromanski and Tal Malkin. The Power of the Dinur-Nissim Algorithm: Breaking Privacy of Statistical and Graph Databases. In *PODS*, pages 65–76. ACM, 2012.

4 Anindya De. Lower Bounds in Differential Privacy. In *TCC*, pages 321–338, 2012.

5 Irit Dinur and Kobbi Nissim. Revealing Information while Preserving Privacy. In *PODS*, pages 202–210. ACM, 2003.

6 Yen Do and Van Vu. The Spectrum of Random Kernel Matrices: Universality Results for Rough and Varying Kernels. *Random Matrices: Theory and Applications*, 2(03), 2013.

[11] This may not be a realistic assumption in many practical scenarios. For example, an individual's salary and postal address code are typically correlated.

7 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC*, volume 3876 of *LNCS*, pages 265–284. Springer, 2006.

8 Cynthia Dwork, Frank McSherry, and Kunal Talwar. The Price of Privacy and the Limits of LP Decoding. In *STOC*, pages 85–94. ACM, 2007.

9 Cynthia Dwork and Sergey Yekhanin. New Efficient Attacks on Statistical Disclosure Control Mechanisms. In *CRYPTO*, pages 469–480. Springer, 2008.

10 Arthur E Hoerl and Robert W Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

11 Prateek Jain and Abhradeep Thakurta. Differentially Private Learning with Kernels. In *ICML*, pages 118–126, 2013.

12 Lei Jia and Shizhong Liao. Accurate Probabilistic Error Bound for Eigenvalues of Kernel Matrix. In *Advances in Machine Learning*, pages 162–175. Springer, 2009.

13 Noureddine El Karoui. The Spectrum of Kernel Random Matrices. *The Annals of Statistics*, pages 1–50, 2010.

14 Shiva Prasad Kasiviswanathan, Mark Rudelson, and Adam Smith. The Power of Linear Reconstruction Attacks. In *SODA*, pages 1415–1433, 2013.

15 Shiva Prasad Kasiviswanathan, Mark Rudelson, Adam Smith, and Jonathan Ullman. The Price of Privately Releasing Contingency Tables and the Spectra of Random Matrices with Correlated Rows. In *STOC*, pages 775–784, 2010.

16 Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the Kernel Matrix with Semidefinite Programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

17 James Mercer. Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, pages 415–446, 1909.

18 Martin M Merener. Polynomial-time Attack on Output Perturbation Sanitizers for Real-valued Databases. *Journal of Privacy and Confidentiality*, 2(2):5, 2011.

19 S. Muthukrishnan and Aleksandar Nikolov. Optimal Private Halfspace Counting via Discrepancy. In *STOC*, pages 1285–1292, 2012.

20 Mark Rudelson. Recent Developments in Non-asymptotic Theory of Random Matrices. *Modern Aspects of Random Matrix Theory*, 72:83, 2014.

21 Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge Regression Learning Algorithm in Dual Variables. In *ICML*, pages 515–521, 1998.

22 Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A Generalized Representer Theorem. In *COLT*, pages 416–426, 2001.

23 Bernhard Scholkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

24 John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

25 John Shawe-Taylor, Christopher KI Williams, Nello Cristianini, and Jaz Kandola. On the Eigenspectrum of the Gram matrix and the Generalization Error of Kernel-PCA. *Information Theory, IEEE Transactions on*, 51(7):2510–2522, 2005.

26 Vikas Sindhwani, Minh Ha Quang, and Aurélie C Lozano. Scalable Matrix-valued Kernel Learning for High-dimensional Nonlinear Multivariate Regression and Granger Causality. *arXiv preprint arXiv:1210.4792*, 2012.

27 Roman Vershynin. Introduction to the Non-asymptotic Analysis of Random Matrices. *arXiv preprint arXiv:1011.3027*, 2010.